同济大学软件学院 2019年春季数据挖掘课程作业2

一、数据集

本次作业采用的数据集主要基于用户在Yelp和大众点评的个人信息及行为痕迹，包括用户的头像数据、用户名、用户评论数据、用户异常行为数据、用户上传图片数据等。两类数据集具体格式如下：

1. **大众点评用户数据: hw2\_data/dianping/**
   1. **用户头像数据**

该数据集包括来自大众点评网站11,948位用户的头像数据，dianping\_avatar.zip文件夹包含所有的头像图片，图片名称为userid\_avatar.jpg，如128884\_avatar.jpg。

* 1. **用户名数据**

该数据集包括来自大众点评网站11,948位用户的用户名。文件命及数据格式如下：

* + dianping\_username.csv

| user\_id | username |
| --- | --- |
| user\_id1 | username1 |
| user\_id2 | username2 |

* 1. **用户评论数据**

该数据集包括来自大众点评网站11,948位用户的评论数据，评论总数为572,773条，评论以文本形式给出。数据集格式如下，comment\_id为评论id，user\_id为发表评论的用户的id，shop\_id为评论对应的商家id，comment\_rank为用户对商家的打分，content为用户评论内容。

* + dianping\_review.csv

| comment\_id | user\_id | shop\_id | comment\_rank | content |
| --- | --- | --- | --- | --- |
| comment\_id1 | user\_id2 | shop\_id1 | comment\_rank1 | content1 |
| commen\_id2 | user\_id3 | shop\_id2 | comment\_rank2 | content2 |

* 1. **异常用户数据**

该数据集包括来自大众点评网站11,948位用户是否为行为异常（例如发表大量恶意评论， 对同一家店短时间内评论态度差异较大等）用户的数据，其中is\_malicious有0和1两种值，0代表用户行为正常， 1代表用户行为异常。文件名以及数据格式如下：

* + dianping\_malicious\_user.csv

| index | user\_id | is\_malicious |
| --- | --- | --- |
| 0 | user\_id1 | label1 |
| 1 | user\_id2 | label2 |

* 1. **用户性别数据**

该数据集为大众点评网站11,948名用户的性别数据，其中1代表female，0代表male，文件名和数据格式如下：

* + dianping\_gender.csv

| user\_id | gender |
| --- | --- |
| user\_id1 | label1 |
| user\_id2 | label2 |

1. **Yelp** 用户数据
   1. **用户名数据**

该数据集包括了来自Yelp网站的223,699位用户的用户名，数据集格式同大众点评用户名数据集格式。文件名如下：

* + yelp\_username.csv
  1. **用户评论数据**

该数据集包括了来自Yelp网站的223,699位用户共计2,865,907条评论数据，评论以文本格式给出，数据集格式如下

* + yelp\_review\_text.csv

| review\_id | user\_id | text |
| --- | --- | --- |
| review\_id1 | user\_id1 | text1 |
| review\_id2 | user\_id2 | text2 |

* 1. **用户消费品类数据**

该数据集包括了来自Yelp网站的223,699位用户消费商家的品类信息，用户消费数据集在第一次作业的 user\_business\_223699.json 数据集中给出，商家品类信息在第一次作业的business\_163665.json 数据集中给出。

* 1. **用户发表图片数据**

该数据集的具体图片数据及信息在第一次作业中的photos\_64048\_user\_719及user\_photo\_719.json 中已给出。

注：在现有数据前提下，如果有需要可以自行进行数据集的扩展。例如根据点评用户的 uid 适量爬取发表的图片，根据yelp用户的uid适量爬取头像等。（扩展与否不影响评分）

二、任务

利用上述大众点评或Yelp的数据集完成对用户的性别预测实验，实验任务如下：

* 1. 分类器：根据百度云中提供的特征，分别采用以下分类器，调整参数进行实验。其中Neural Network为可选项，其他三项为必选项：
     1. Logistic Regression
     2. SVM
     3. Random Forest Classifier
     4. (Optional, BONUS) Neural Network (ReLu, Softmax)

在实验过程中需要将数据集进行划分，采用训练集，验证集和测试集的划分方式进行分类实验。

* 1. 实验结果评估：对所用到的每个分类器，计算其在测试集上的 Accuracy, Precision, Recall, F1-Score值。
  2. （附加题，有额外加分）提供的数据集为用户性别预测提供了多种基本的特征。在这些基本特征的前提下，自由选取特征进行处理或者组合，形成新的特征。利用新提取的特征重复任务1、2中的实验。选取基本特征时要适当考虑所选取特征的含义，例如malicious标签对用户性别预测是否有帮助。对选取的特征也可以进行进一步的处理，例如对用户评论进行情感分析，词量统计或者主题提取；对用户购买商品的类别进行划分等；对用户发表的图片进行处理；将用户名转化为词向量等。
  3. 综合以上实验，分析结果并完成实验报告。

注：

* 1. ~~本次实验结果涉及到的所有数据集均可在~~[**~~http://10.60.43.58:9579/hw2\_data/~~**~~上进行下载~~](http://10.60.43.58:9579/hw2_data/上进行下载)由于服务器所在机房周末断电，目前服务器正在修复且不是很稳定， 请大家从指定的百度云盘中下载数据。
  2. 提供的用户数据集是随机选取的，可根据实际情况进行筛选和处理，实验中不一定要用到全量数据。
  3. 作业中介绍的特征不一定要全部用到，可根据实验结果进行选择。
  4. 分类器中的**Neural Network**实现为可选项，不完成不影响评分，完成可有额外加分。
  5. 由于Yelp数据集没有提供标签，如果选择该数据集需要自行解决标签问题，并且有额外加分。

三、提交

提交日期：**2019-4-27 23:59:59**，提交至Piazza。提交内容要求：

提交文件命名为学号\_姓名(中文)\_hw2.zip。压缩包中包含以下内容：

* 1. 源代码文件。
  2. README文件，介绍运行环境和运行方式。
  3. 实验报告文件，包括数据预处理、特征提取、特征选取，分类器参数、实验结果、相应的图表以及对实验结果的比较分析等。
  4. 实验结果文件。大众点评的预测结果或yelp的预测结果各一个文件，均为csv文件格式。每行具体内容如下：

Dianping: [ user\_id, predict\_label, true\_label ]

Yelp: [ user\_id, predict\_label, true\_label ]

如果完成了附加题， 请将附加题的预测结果和任务1的预测结果区分开，文件格式和上述格式相同。